Forecasting Urban Air Quality for Public Health Advisories - Final Report

Motivation:
Air monitoring is crucial for human well-being, especially as major Indian cities face significant public health risks from severe air pollution events. Delhi is a highly pollution-challenged megacity where the Air Quality Index (AQI) frequently goes above 300-400 (classified as "Poor" or "Very Poor") during the winter and post-monsoon seasons.

Problem Statement:
Air pollution poses a major health and environmental challenge in India. Cities such as Delhi frequently suffer from hazardous air quality episodes driven by traffic, industrial emissions, meteorology, biomass burning, and seasonal variations. Despite the urgency, citizens and authorities lack an early-warning system capable of forecasting AQI levels in advance. A reliable short-term AQI forecasting tool can support:

- Public health advisories
- Policy decisions and traffic regulation
- Pollution control operations
- Personal planning for sensitive groups This project aims to address this gap through data-driven forecasting.

Datasets:
The dataset is from kaggle and it captures air-quality measurements across multiple Indian cities (around 26 major cities) between 2015 and 2020.
(Source: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india?resource=download)
Data are recorded at hourly and daily intervals, depending on the station.
Variables include common air-pollutants and indices such as:

- Particulate matter: PM2.5 and PM10
- Gaseous pollutants / other pollutants — e.g. NO2, SO2, CO, O3 (and possibly others like NO, NOx, NH_3, volatile organics depending on station).
- An overall Air Quality Index (AQI) value computed per record.
- A categorical label (sometimes called “AQI_Bucket”) that classifies air quality into categories like “good”, “satisfactory”, “moderate”, “poor”, “very poor”, “severe”

We have considered dataset from Kaggle where it consists 5 different dataset files total of 296.55 MB

- **city_day.csv**
- **city_hour.csv**
- station_day.csv
- station_hour.csv
- stations.csv

Methodology:

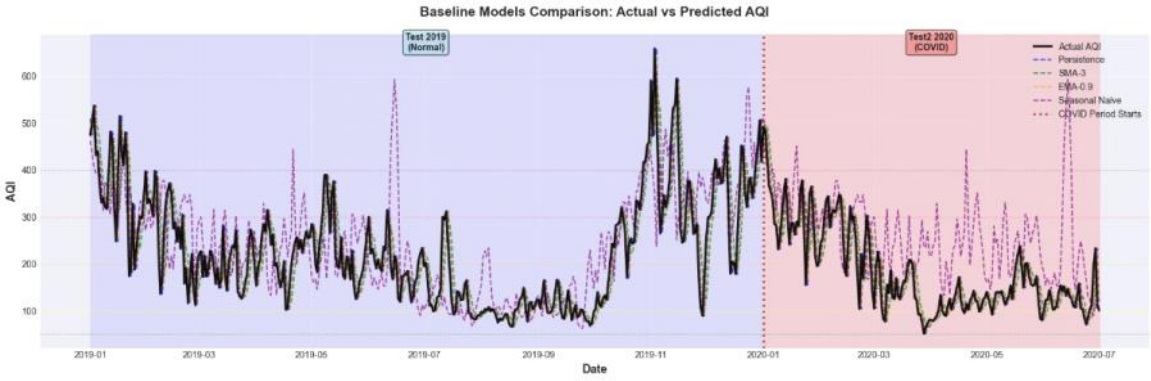
1. Data Preprocessing
   - Resampling hourly data to daily averages
   - Handling missing values (interpolation / forward fill)
   - Smoothing and noise reduction
   - Feature engineering (lags, rolling averages, seasonal indicators)
2. Baseline Models
   - Simple Moving Average (SMA): A naïve baseline that smooths the series using the average of previous N days.
   - Exponential Moving Average (EMA): A faster-responding baseline giving higher weight to recent observations. These serve as reference points to measure the improvement of advanced models.
3. Advanced Forecasting Models
   - SARIMA : Captures seasonal patterns, models trend + seasonality + autocorrelation. Strong for classical time-series with periodic behavior.
   - LSTM (Long Short-Term Memory) : Deep-learning model for sequence data. Learns long-term dependencies and non-linear patterns, handling complex relationships between pollutants and meteorology.
   - XGBoost (Extreme Gradient Boosting) : State-of-the-art tree-based model. Excellent for tabular time-series with engineered lag features, handling missing data and non-linear interactions effectively. Often outperforms neural networks in real-world AQI forecasting.
4. Evaluation Metrics The models are evaluated using standard regression and categorical metrics:
   - MAE (Mean Absolute Error)
   - RMSE (Root Mean Square Error)
   - MAPE (Mean Absolute Percentage Error)
   - R^2 Score
   - AQI Category Accuracy (Correct classification into Good / Moderate / Unhealthy etc.)

Key Results and Insights:

1. Next 24–72 hour AQI forecasts
2. Visualization plots for actual vs predicted values
3. Performance metrics for all models
4. Best performing model recommendation
5. GitHub link: https://github.com/prakashsintel/Forecasting-Urban-Air-Quality-

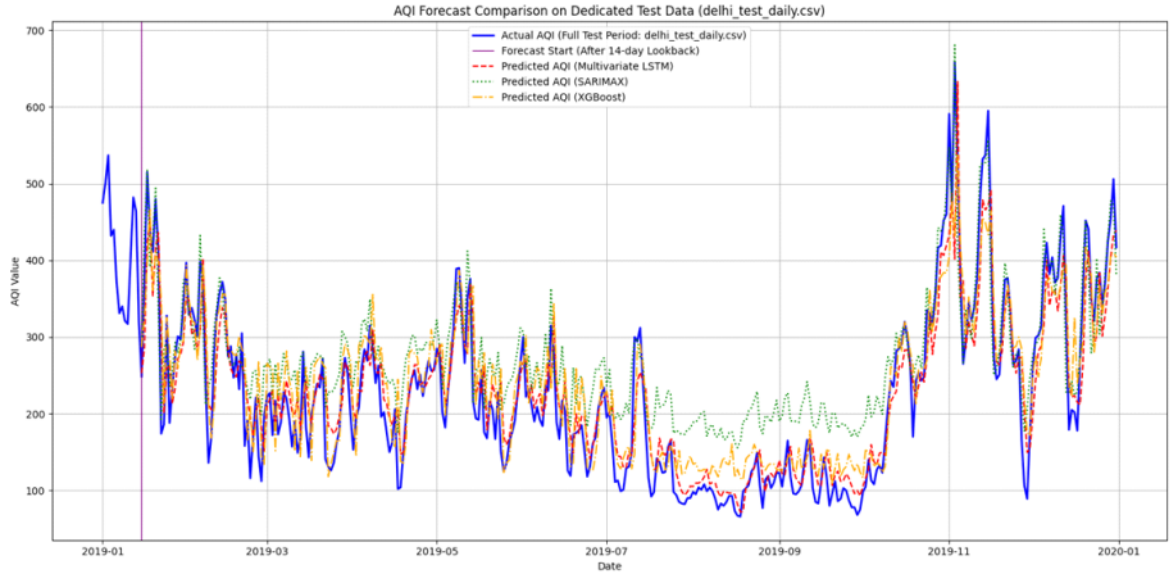
NAÏVE Model Results

Model	MAE	RMSE	MAPE	R2	Category_Accuracy
Persistence	38.48	53.99	17.72%	0.79	64.66%
EMA-0.9	38.87	54.38	18.01%	0.79	64.38%
SMA-3	48.49	66.06	23.09%	0.68	55.07%
Seasonal_Naive-365	70.17	95.05	37.05%	0.34	41.64%



Advanced model Results

Model	MAE	MAPE	RMSE	R2	Category Accuracy
SARIMAX	54.84	38.73%	63.06	0.6899	47.01%
LSTM	32.90	17.14%	44.07	0.8485	68.38%
XGBoost	39.65	22.96%	51.78	0.7909	59.83%



Limitation and possible future improvements

1. Work on more recent AQI data.

Contribution from each team member

1. Pothukanuri Sai Venkat - Data cleaning and data processing.
2. Shivam Kumar - Complete EDA and Naïve model based prediction.
3. Prakash S - Data modelling (SARIMAX and XGBOOST) and evaluation.
4. Sachin R - Data modelling (LSTM) and evaluation.
5. Everyone contributed equally for documentation and Reports and git-hub deployment.